Variational Combinatorial Sequential Monte Carlo for Bayesian Phylogenetic Inference

Antonio Moretti, Liyi Zhang & Itsik Pe'er
Department of Computer Science
Columbia University
New York, NY 10027, USA
amoretti@cs.columbia.edu

Abstract

Bayesian phylogenetic inference is often conducted via local or sequential search algorithms such as random-walk Markov chain Monte Carlo or Combinatorial Sequential Monte Carlo. These methods sample tree topologies and branch lengths, however when performing evolutionary parameter learning, they require long runs with inefficient state space exploration. Here we introduce Variational Combinatorial Sequential Monte Carlo (VCSMC), a novel Variational Inference method that simultaneously performs both parameter inference and model learning. VCSMC uses sequential search to construct a variational objective defined on the composite space of phylogenetic trees. We show that VCSMC is computationally efficient and explores higher probability spaces when compared with state-of-the-art Hamiltonian Monte Carlo methods.

1 Introduction

Bayesian phylogenetic inference plays a central role in molecular evolutionary biology due to its ability to represent evolutionary uncertainty and incorporate prior information. Inference often involves three distinct tasks: (i) sampling from a discrete distribution to approximate an intractable summation over tree topologies, (ii) for each tree, integrating over the continuous parameters and branch lengths that govern the evolutionary model of interest, and (iii) performing parameter estimation or model learning. The sampling of tree topologies and branch lengths is typically accomplished via local search algorithms such as random-walk Markov chain Monte Carlo (Huelsenbeck & Ronquist, 2001) or sequential search algorithms such as Combinatorial Sequential Monte Carlo (Bouchard-Côté et al., 2012). Sophisticated proposal methods based on Hamiltonian Monte Carlo or particle MCMC have been suggested to sample from composite spaces and infer evolutionary parameters (Dinh et al., 2017; Wang et al., 2015; Wang & Wang, 2020), however these methods are often difficult to implement, slow to converge and heavily dependent upon heuristics.

Variational Inference (VI) is a computationally efficient alternative to MCMC that simultaneously performs both inference and model learning. VI posits an approximate distribution and then recovers parameters of both the model and approximation by maximizing a lower bound to the log marginal likelihood. One approach to learning variational distributions on phylogenetic trees is to parameterize a tree as a sequence of *subsplits*, or ordered partitions on clades (Zhang & Matsen IV, 2018) and to recast the problem as a Bayesian network. One drawback of this setup is that the support of the conditional probability tables scales exponentially with the number of taxa (Zhang & Matsen IV, 2019). A body of recent work has established connections between VI and sequential search by defining a variational family of distributions on hidden Markov models, where Sequential Monte Carlo is used as the marginal likelihood estimator (Le et al., 2018; Naesseth et al., 2018; Moretti et al., 2019c; 2020). Here we introduce Variational Combinatorial Sequential Monte Carlo (VCSMC), a novel variational objective and structured approximate posterior defined on the composite space of phylogenetic trees. Unlike standard variational SMC methods, our objective is constructed from *partial* states where the likelihood is not directly available and where states are formed by sampling from a large combinatorial set. VCSMC provides suitable estimates of the posterior when applied to a benchmark dataset of primate mitochondrial DNA and performs favorably when compared with the state of the art HMC methods.

2 Background

Phylogenetic Trees We wish to infer a latent bifurcating tree that describes the evolutionary relationships among a set of observed molecular sequences. A phylogeny is defined by a tree topology τ and a set of branch lengths $\mathcal{B}$. A *tree topology* is defined as a connected acyclic graph (V, E) where V is a set of vertices and E is a set of edges. *Leaf nodes* denote vertices of degree 1 and correspond to observed taxa. *Internal nodes* designate vertices of degree 3 (one parent and two children) and represent unobserved taxa (e.g. DNA bases of ancestral species). A special vertex called the *root node* of degree 2 (two children) represents the common evolutionary ancestor of all taxa.

For each edge $e \in E$, we associate a *branch length*, denoted $b(e) \in \mathbb{R}_{>0}$, $b(e) \in \mathcal{B}$. The branch length captures the intensity of the evolutionary changes between two vertices. An *ultrametric tree* is one with constant evolutionary rate along all paths from v to its descendants. *Nonclock trees* are general trees that do not require ultrametric assumptions. In this work we focus on phylogenetic inference methods for nonclock trees as these are most pertinent to biologists.

Bayesian Phylogenetic Inference Let $\mathbf{Y} = \{Y_1, \cdots, Y_M\} \in \Omega^{NxM}$ denote the observed molecular sequences with characters in Ω of length M over N species. Bayesian inference requires specifying the prior density and likelihood function over tree topology τ, branch length set $\mathcal{B}$ and generative model parameters θ to write the joint posterior,

$$P(\mathcal{B}, \tau, \theta | \mathbf{Y}) = \frac{P(\mathbf{Y}|\tau, \mathcal{B}, \theta) P(\tau, \mathcal{B}|\theta) P(\theta)}{P(\mathbf{Y})}. \tag{1}$$

The prior is uniform over topologies and a product of independent exponential distributions over branch lengths with rate λ_{bl}. The evolution of each site is modeled independently using a continuous time Markov chain with rate matrix $\mathbf{Q}$. Let $\zeta_{v,m}$ denote the state of genome for species v at site m and define the evolutionary model along branch $b(v \to v')$:

$$P(\zeta_{v',s} = j | \zeta_{v,s} = i) = \exp\left(b(e)\mathbf{Q}_{i,j}\right). \tag{2}$$

The likelihood of a given phylogeny $P(\mathbf{Y}|\tau, \mathcal{B}, \theta) = \prod_{i=1}^{M} P(Y_i|\tau, \mathcal{B}, \theta)$ can be evaluated in linear time using the sum-product or pruning algorithm (Felsenstein, 1981), however the normalization constant $P(\mathbf{Y})$ requires marginalizing the $(2N-3)!!$ distinct topologies (Semple & Steel, 2003) which is intractable.

Combinatorial Sequential Monte Carlo CSMC is a method to sample from a probability measure $\bar{\pi}$ by performing inference on a sequence of increasing probability spaces (Wang et al., 2015). The target measure $\bar{\pi}$ and its normalization constant $\|\pi\|$ corresponding to the numerator and denominator in Eq. (1) are approximated by sequential importance resampling in R steps. Unlike standard SMC methods, the target is defined on a combinatorial set (the space of tree topologies $\mathcal{T}$). K sampled *partial states* (or *particles*) $\{s_{r,k}\}_{k=1}^{K} \in \mathcal{S}_r$ are drawn at each rank r and used to form a discrete positive measure,

$$\pi_{r,k} = \|\pi_{r-1,k}\| \frac{1}{K} \sum_{k=1}^{K} w_{r,k} \delta_{s,k}(s) \qquad \forall s \in \mathcal{S}, \tag{3}$$

where δ_s is the Kronecker delta and $w_{r,k}$ are the importance weights. Resampling ensures that particles remain on areas of high probability mass. Each resampled state $\tilde{s}_{r-1,k}$ of rank $r-1$ is then extended to a state of rank r by drawing from a proposal distribution $s_{r,l} \sim \nu^{+}_{s_{r,k}} : \mathcal{S} \to [0,1]$. The importance weights are computed as follows:

$$w_{r,k} = w(\tilde{s}_{r-1,k}, s_{r,k}) = \frac{\pi(s_{r,k})}{\pi(\tilde{s}_{r-1,k})} \cdot \frac{\nu^{-}_{s_{r,k}}(\tilde{s}_{r-1,k})}{\nu^{+}_{\tilde{s}_{r,k}}(s_{r,k})}, \tag{4}$$

where $\nu^{-}_{s_{r,k}}$ is a probability density over $\mathcal{S}$ correcting an over-counting problem (Wang et al., 2015). The procedure is summarized in Algorithm 1 of the Appendix. An unbiased estimate for the marginal likelihood can be constructed from the weights which converges in L^2 norm,

$$\hat{\mathcal{Z}}_{CSMC} := \|\pi_{R,K}\| = \prod_{r=1}^{R} \left(\frac{1}{K} \sum_{k=1}^{K} w_{r,k} \right) \to \|\pi\|. \tag{5}$$

Variational Inference VI is a technique for approximating the posterior $\log P_\theta(\mathcal{B}, \tau|\mathbf{Y})$ when marginalization of latent variables is not analytically feasible. By introducing a tractable distribution $Q_\phi(\mathcal{B}, \tau|\mathbf{Y})$ it is possible to form a lower bound to the log-likelihood:

$$\log P_\theta(\mathbf{Y}) \geq \mathcal{L}_{\text{ELBO}}(\theta, \phi, \mathbf{Y}) := \mathop{\mathbb{E}}_{Q}\left[\log \frac{P_\theta(\mathbf{Y}, \mathcal{B}, \tau)}{Q_\phi(\mathcal{B}, \tau|\mathbf{Y})}\right]. \tag{6}$$

Auto Encoding Variational Bayes (Kingma & Welling, 2013) (AEVB) simultaneously trains $Q_\phi(\mathcal{B}, \tau|\mathbf{Y})$ and $P_\theta(\mathbf{Y}, \mathbf{Z})$. The expectation in Eq. (6) is approximated by averaging Monte Carlo samples from $Q_\phi(\mathcal{B}, \tau|\mathbf{Y})$ which are reparameterized by evaluating a deterministic function of a ϕ-independent random variable.

3 Variational Combinatorial Sequential Monte Carlo

Variational Objective The idea of VCSMC is to simultaneously train the target and proposal distribution by maximizing a lower bound to the data log-likelihood, while using CSMC as the marginal likelihood estimator. We begin by defining a structured approximate posterior which factorizes over rank events. To do so, we will change notation from CSMC writing the resampled state $\tilde{s}_{r-1,k}$ as $s_{r-1}^{a_{r-1}^k}$ to make explicit the dependency of $\tilde{s}_{r-1}$ on its resampled index a_{r-1}^k. Let $q_\phi(s_{r,k}|s_{r-1}^{a_{r-1}^k})$ denote conditional the probability of state $s_{r,k}$ given the resampled state at the previous rank $s_{r-1}^{a_{r-1}^k}$. Subscripts ϕ and ψ denote discrete and continuous proposal parameters respectively:

$$\begin{aligned} Q_{\phi,\psi}\left(\mathcal{S}_{1:R}^{1:K}\right) :=& \left(\prod_{k=1}^{K} q_\phi(s_{1,k}) \cdot q_\psi(\mathcal{B}_{1,k})\right) \\ &\times \left(\prod_{k=1}^{K}\prod_{r=1}^{N-1} q_\phi\left(s_{r,k}|s_{r-1}^{a_{r-1}^k}\right) \cdot q_\psi\left(\mathcal{B}_{r,k}|\mathcal{B}_{r-1}^{a_{r-1}^k}\right) \cdot \textsc{Categorical}\left(a_{r-1}^k|\bar{w}_{r-1}^{1:K}\right)\right). \end{aligned} \tag{7}$$

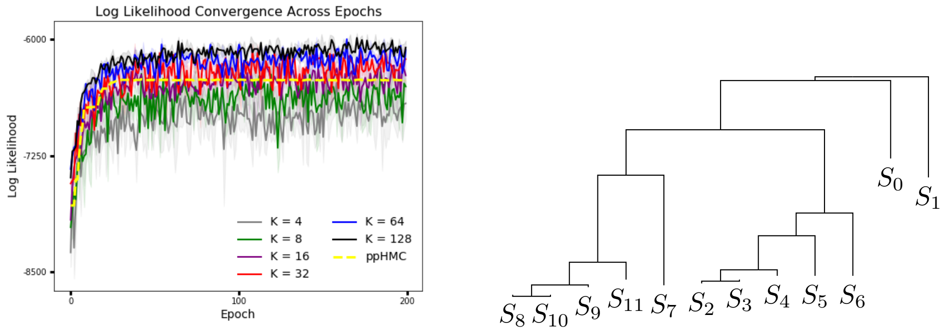
At the final rank event, an unbiased approximation to the likelihood is formed by averaging over importance weights, which, in turn represent the sample phylogenies that are constructed iterativly. A multi-sample variational objective formed is via the lower bound:

$$\mathcal{L}_{VCSMC} := \mathop{\mathbb{E}}_{Q}\left[\log \hat{\mathcal{Z}}_{VCSMC}\right], \qquad \hat{\mathcal{Z}}_{VCSMC} := \|\pi_{R,K}\| = \prod_{r=1}^{R}\left(\frac{1}{K}\sum_{k=1}^{K} w_{r,k}\right) \tag{8}$$

The presence of the DISCRETE densities over partial states presents a challenge for variational reparameterization. Unlike standard variational SMC methods, states are formed by sampling from a large combinatorial set. We take two approaches, the first is to drop discrete terms from the gradient estimates. The second is to reparameterize these terms as Gumbel-Softmax random variables forming a differentiable approximation through a convex relaxation over the simplex. Continuous proposal terms are drawn by evaluating a deterministic function of a ψ-independent random variable.

Implementation Details Constructing the objective $\mathcal{L}_{VCSMC}$ is done iteratively in three steps. The EXTENDPARTIALSTATE procedure requires selecting two partial states to coalesce by sampling without replacement. This is accomplished by defining Gumbel-Softmax random variables. The uniform log-probability for each index is perturbed by adding independent Gumbel distributed noise, after which the largest two elements are returned. For example let $U \sim \textsc{Uniform}(0, 1)$, we then form $G = \gamma - \log(-\log U)$ so that G can be reparameterized as $G' = G + \gamma$. The RESAMPLE procedure can also be reparameterized similarly by defining Gumbel-Softmax random variables.

The COMPUTEWEIGHTS step requires some care. In order to compute importance weights, the likelihood of a partial state must be computed using the sum-product algorithm, however the probability measure π is only defined on the target space of trees $\mathcal{T}$, and not the larger sample space of partial states $\mathcal{S} := \cup_r \mathcal{S}_r$. Intuitively, the sum-product or pruning algorithm yields a maximum likelihood estimate for an evolutionary tree, but partial states contain disjoint subtrees or disjoint leaf nodes. To illustrate this, consider the jump chain for the partial state $\{A, B\}$ defined on the four taxa $\{A, B, C, D\}$ written as $s_1 = \{\{A, B\}, \{C\}, \{D\}\}$. This partial state admits three possible



(a) Log likelihood across epochs

(b) Phylogeny sampled from the posterior

Figure 1: (Left): Log likelihood values for K = $\{4, 8, 16, 32, 64, 128\}$ samples of VCSMC on the primates data averaged across 3 random seeds. Higher values of K produce tighter ELBO / larger log likelihood values with lower stochastic gadient noise. VCSMC with $K \geq 16$ outperforms probabilistic path Hamiltonain Monte Carlo (ppHMC) which is shown (yellow) for comparison. (Right): A single nonclock phylogeny sampled from the posterior with probability proportional to the importance weights at the final step. From left to right: M Mulatta, M Sylvanus, M Fascicularis, Saimiri Sciureus, Macaca Fuscata, Homo Sapiens, Pan, Gorilla, Pongo, Hylobates, Tarsius Syrichta, Lemur Catta. The leftmost clade partitions monkeys whereas the central and right clades partition hominids and prosimians respectively.

evolutionary trees (depicted in Fig 2 of the Appendix). The likelihood for each of these phylogenies contains a factor corresponding to the message passed from $\{A, B\}$ to the parent node $\text{PA}(A, B)$. At the root node, in order to form the likelihood from a distribution over discrete characters, the pruning algorithm evaluates the inner product of PA and the prior η (the stationary state of $\mathbf{Q}$). One extension of the target measure π into a measure on $\mathcal{S}$ suggested by (Wang et al., 2015) is to treat all elements of the jump chain as trees (in this case, the subtree consisting of $\{A, B\}$ or $\text{PA}(A, B)$ and non-coalescing singletons $\{C\}$ and $\{D\}$). The contribution of each of the elements in the jump chain to the likelihood is multiplied by taking the inner product of each distribution over characters with η. This extension has the advantage of passing information from the non-coalescing elements to the local weight update. We explore other extensions in future work.

4 Results

Primate Mitochondrial DNA We evaluate VCSMC on a benchmark dataset of nucleotide sequences of homologous fragments of primate mitochondrial DNA (Hayasaka et al., 1988). The dataset consists of 12 taxa $\{S_0, \cdots, S_{11}\}$ over 898 sites admitting 13,749,310,575 distinct tree topologies. The set of taxa includes five species of homonoids, four species of old world monkeys, one species of new world monkey and two species of prosimians. VCSMC is run with $K = \{4, 8, 16, 32, 64, 128\}$ particles, averaged over 3 random seeds. Fig 1 (left) shows higher values of K produce larger log likelihood values (tighter ELBO values) with lower stochastic gradient noise. VCSMC with $K \geq 16$ outperforms probabilistic path Hamiltonain Monte Carlo (ppHMC) shown (yellow trace) for comparison. Fig 1 (right) illustrates a single phylogeny sampled from the posterior with probability proportional to the importance weights at the final step. From left to right: M Mulatta, M Sylvanus, M Fascicularis, Saimiri Sciureus, Macaca Fuscata, Homo Sapiens, Pan, Gorilla, Pongo, Hylobates, Tarsius Syrichta, Lemur Catta. The leftmost clade partitions monkeys whereas the central and right clades partition hominids and prosimians respectively.

5 Conclusion

We have sketched VCSMC, a method for model inference and parameter learning in Bayesian phylogenetics. To our knowledge, VCSMC is the first method to define a variational objective on the composite space of phylogenetic trees using Sequential Monte Carlo. VCSMC is written in Tensorflow. An implementation is available online at https://github.com/amoretti86/phylo.

References

Alexandre Bouchard-Côté, Sriram Sankararaman, and Michael Jordan. Phylogenetic inference via sequential monte carlo. *Systematic biology*, 61:579–93, 01 2012. doi: 10.1093/sysbio/syr131.

J. Rodney Brister, Danso Ako-adjei, Yiming Bao, and Olga Blinkova. NCBI Viral Genomes Resource. *Nucleic Acids Research*, 43(D1):D571–D577, 11 2014. ISSN 0305-1048. doi: 10.1093/nar/gku1207. URL https://doi.org/10.1093/nar/gku1207.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders, 2015.

Vu Dinh, Arman Bilge, Cheng Zhang, and Frederick A. Matsen, IV. Probabilistic path Hamiltonian Monte Carlo. volume 70 of *Proceedings of Machine Learning Research*, pp. 1009–1018, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/dinh17a.html.

J Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981. URL http://www.ncbi.nlm.nih.gov/pubmed/7288891?ordinalpos=5&itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_DefaultReportPanel.Pubmed_RVDocSum.

K Hayasaka, T Gojobori, and S Horai. Molecular phylogeny and evolution of primate mitochondrial DNA. *Molecular Biology and Evolution*, 5(6):626–644, 11 1988. ISSN 0737-4038. doi: 10.1093/oxfordjournals.molbev.a040524. URL https://doi.org/10.1093/oxfordjournals.molbev.a040524.

Daniel Hernandez, Antonio Moretti, Ziqiang Wei, S. Saxena, John Cunningham, and Liam Paninski. A novel variational family for hidden nonlinear markov models. *CoRR*, abs/1811.02459, 2018a.

Daniel Hernandez, Antonio Khalil Moretti, Ziqiang Wei, Shreya Saxena, John Cunningham, and Liam Paninski. Nonlinear evolution via spatially-dependent linear dynamics for electrophysiology and calcium data. *Neurons, Behavior, Data analysis and Theory*, 2018b.

John P. Huelsenbeck and Fredrik Ronquist. MRBAYES: Bayesian inference of phylogenetic trees . *Bioinformatics*, 17(8):754–755, 08 2001. ISSN 1367-4803. doi: 10.1093/bioinformatics/17.8.754. URL https://doi.org/10.1093/bioinformatics/17.8.754.

Sebastian Höhna and Alexei Drummond. Guided tree topology proposals for bayesian phylogenetic inference. *Systematic biology*, 61:1–11, 01 2012. doi: 10.1093/sysbio/syr074.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.

Clemens Lakner, Paul van der Mark, John P. Huelsenbeck, Bret Larget, and Fredrik Ronquist. Efficiency of Markov Chain Monte Carlo Tree Proposals in Bayesian Phylogenetics. *Systematic Biology*, 57(1):86–103, 02 2008. ISSN 1063-5157. doi: 10.1080/10635150801886156. URL https://doi.org/10.1080/10635150801886156.

Tuan Anh Le, Maximilian Igl, Tom Rainforth, Tom Jin, and Frank Wood. Auto-encoding sequential monte carlo. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJ8c3f-0b.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables, 2016.

Antonio Moretti, Andrew Stirn, Gabriel Marks, and Itsik Pe'er. Autoencoding topographic factors. *Journal of Computational Biology*, 26(6):546–560, 2019a.

Antonio K Moretti, Zizhao Wang, Luhuan Wu, and Itsik Pe'er. Smoothing nonlinear variational objectives with sequential monte carlo. *ICLR Workshops*, 2019b. URL https://openreview.net/pdf?id=HJg24U8tuE.

Antonio Khalil Moretti, Zizhao Wang, Luhuan Wu, Iddo Drori, and Itsik Pe'er. Particle smoothing variational objectives. *CoRR*, abs/1909.09734, 2019c.

Antonio Khalil Moretti, Zizhao Wang, Luhuan Wu, Iddo Drori, and Itsik Pe'er. Variational objectives for markovian dynamics with backward simulation. *European Conference on Artificial Intelligence*, 2020.

D.A. Morrison. Multiple sequence alignment for phylogenetic purposes. *Aust. Syst. Bot.*, 19:476–539, 01 2006.

Christian Naesseth, Scott Linderman, Rajesh Ranganath, and David Blei. Variational sequential monte carlo. volume 84 of *Proceedings of Machine Learning Research*, pp. 968–977, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR. URL http://proceedings.mlr.press/v84/naesseth18a.html.

Fredrik Ronquist, Maxim Teslenko, Paul Mark, Daniel Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc Suchard, and John Huelsenbeck. Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, 61: 539–42, 03 2012. doi: 10.1093/sysbio/sys029.

Charles Semple and Mike Steel. Phylogenetics. 2003.

Liangliang Wang, Alexandre Bouchard-Côté, and Arnaud Doucet. Bayesian phylogenetic inference using a combinatorial sequential monte carlo method. *Journal of the American Statistical Association*, 01 2015. doi: 10.6084/M9.FIGSHARE.1478005.

Shijia Wang and Liangliang Wang. Particle gibbs sampling for bayesian phylogenetic inference, 2020.

Cheng Zhang and Frederick A Matsen IV. Generalizing tree probability estimation via bayesian networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 1444–1453. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7418-generalizing-tree-probability-estimation-via-bayesian-networks.pdf.

Cheng Zhang and Frederick A Matsen IV. Variational bayesian phylogenetic inference. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SJVmjjR9FX.

Appendix: Additional Details of the Implementation

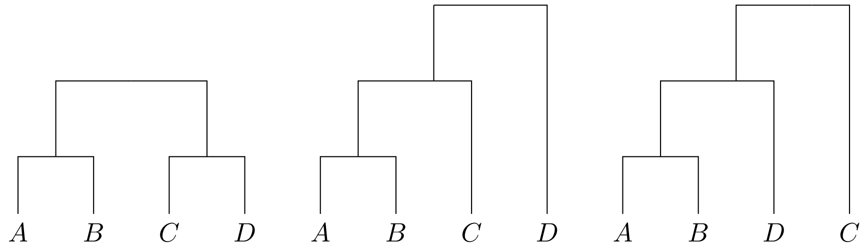


Figure 2: An example of the partial state $s = \{A, B\}$ for four taxa $\{A, B, C, D\}$ illustrated using its dual representation $\mathcal{D}(s)$. The dual state $\mathcal{D}(s) \subseteq \mathcal{T}$ corresponds to the three complete tree topologies. (**left**): $\{\{A, B\}, \{C, D\}\}$ (**center**): $\{\{A, B\}, \{A, B, C\}\}$ and (**right**): $\{\{A, B\}, \{A, B, D\}\}$.

Theorem 1 *(Gershgorin) Let A be an $n \times n$ matrix with entries in $\mathbb{C}$. For each i, let D_i be the disc,*

$$D_i = \left\{ z \in \mathbb{C} : |z - A_{ii}| \leq \sum_{j \neq i} |A_{ij}| \right\} , \tag{9}$$

then the eigenvalues of A lie in $D_1 \cup D_2 \cup \cdots \cup D_n$. It follows that an upper bound on the maximum absolute value for the eigenvalues of A is given by:

$$\max_i \lambda_i \leq \max_i \sum_j |a_{ij}| \, . \tag{10}$$

The likelihood of a given phylogeny is independent across sites and can be evaluated using the sum-product algorithm via the formula:

$$P(\mathbf{Y}|\tau, \mathcal{B}, \theta) := \prod_{i=1}^{M} \sum_{a^i} \eta(a^i_\rho) \prod_{(u,v) \in E(\tau)} \exp\left(-b_{u,v} \mathbf{Q}_{a^i_u, a^i_v}\right) , \tag{11}$$

where ρ is the root node, a^i_u is the assigned character of node u, $E(\tau)$ represents the set of edges in τ and η is the prior or stationary distribution of the Markov chain.

In the experiments, the trainable parameters consist of the components of $\mathbf{Q}_{ij}$ and the branch length distribution rates λ_{bl} for each $q_\psi\left(\mathcal{B}_{r,k}|\mathcal{B}_{r-1}^{a^k_{r-1}}\right)$. The softmax activation is used to constrain and

Algorithm 1: Combinatorial Sequential Monte Carlo

```
0. Initalization. $\forall\, k$, $s_{0,k} \leftarrow \perp$, $w_{0,k} \leftarrow 1/K$;
1. for r = 0 to $|X| - 1$ do
    2. for k=1 to K do
        a. Resample partial states
              $\tilde{s}_{r-1,1}, \cdots, \tilde{s}_{r-1,k} \sim \bar{\pi}_{r-1,k}$
        b. Extend partial states
              $s_{r,k} \sim \nu^{+}_{\tilde{s}_{r-1,k}}$
        c. Compute weights for new particles
              $w_{r,k} = w(\tilde{s}_{r-1,k}, s_{r,k}) = \frac{\pi(s_{r,k})}{\pi(\tilde{s}_{r-1,k})} \cdot \frac{\nu^{-}_{s_{r,k}}(\tilde{s}_{r-1,k})}{\nu^{+}_{\tilde{s}_{r,k}}(s_{r,k})}$
    end
end
```

normalize components so that $\sum_{j \neq i} \mathbf{Q}_{ij} = 1$, ensuring a bound on eigenvalues via the Gershgorin circle theorem. The components of $\mathbf{Q}_{ij}$ and λ_{bl} for each $q_\psi \left(\mathcal{B}_{r,k} | \mathcal{B}_{r-1}^{a_{r-1}^k} \right)$ can also be parameterized as a deep generative model using the output of neural networks. In this setup, the evolution of each site is modeled as a nonlinear function of spatial position on the genome.